Analytica Chimica Acta 702 (2011) 172–177

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Blind image analysis for the compositional and structural characterization of plant cell walls

Pradeep N. Perera[a,*], Martin Schmidt[a], P. James Schuck[b], Paul D. Adams[c,d]

[a] Energy Biosciences Institute, University of California, Berkeley, CA 94720, USA
[b] Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[c] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[d] Department of Bioengineering, University of California Berkeley, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 8 February 2011
Received in revised form 5 June 2011
Accepted 10 June 2011
Available online 18 July 2011

Key words:
Hyperspectral Raman imaging
Image analysis
Entropy minimization
Lignin
Curve resolution
Biomass

ABSTRACT

A new image analysis strategy is introduced to determine the composition and the structural characteristics of plant cell walls by combining Raman microspectroscopy and unsupervised data mining methods. The proposed method consists of three main steps: spectral preprocessing, spatial clustering of the image and finally estimation of spectral profiles of pure components and their weights. Point spectra of Raman maps of cell walls were preprocessed to remove noise and fluorescence contributions and compressed with PCA. Processed spectra were then subjected to *k*-means clustering to identify spatial segregations in the images. Cell wall images were reconstructed with cluster identities and each cluster was represented by the average spectrum of all the pixels in the cluster. Pure components spectra were estimated by spectral entropy minimization criteria with simulated annealing optimization. Two pure spectral estimates that represent lignin and carbohydrates were recovered and their spatial distributions were calculated. Our approach partitioned the cell walls into many sublayers, based on their composition, thus enabling composition analysis at subcellular levels. It also overcame the well known problem that native lignin spectra in lignocellulosics have high spectral overlap with contributions from cellulose and hemicelluloses, thus opening up new avenues for microanalyses of monolignol composition of native lignin and carbohydrates without chemical or mechanical extraction of the cell wall materials.

1. Introduction

The plant cell wall is a complex composite of an interconnected network of cellulose cross-linked by hemicelluloses. Lignin, a phenolic polymer, is also present in many secondary cell walls, providing additional structural support and recalcitrance to degradation. Interest in harvesting cellulose for biofuels has renewed the focus on structural and compositional details of plant cell walls. The natural recalcitrance of lignocellulosics, however, has been a major obstacle to efficient conversion of cellulose into fermentable sugars [1,2]. New analytical tools can play an important role, by gaining knowledge of chemical, structural and spatial insights of plant cell walls, to facilitate the advancement of successful biomass deconstruction as well as to improve our basic knowledge of plant cell walls.

Raman microspectroscopy has been successfully employed to probe many important structural and compositional aspects of the plant cell wall [3–9]. However, current spectral image analysis and reconstruction methods have been limited to peak intensity or peak integration with few exceptions [10], thus making suboptimal use of the multiplexing nature of Raman spectroscopy. In fact, the detailed characterization requires implementation of vigorous multivariate methods to untangle the chemically heterogeneous and structurally complex plant cell wall. In this article, we demonstrate a new non-supervised image analysis strategy which in combination with Raman (or other spectroscopies) can be employed to elucidate chemical, structural and spatial information of the plant cell wall.

Spectral images of plant cell walls present a stern challenge to multivariate data mining methods for the following reasons:

(1) A multi-component system is present at any given image pixel spectrum and often the exact number of components is unknown.
(2) Pure components spectra and their weight are unknown. The spectrum of an isolated component (e.g. lignin) would be different from that in the native environment due to loss of hydrogen

* Corresponding author at: Energy Biosciences Institute, 130, Calvin Laboratory, MC 5230, University of California, Berkeley, CA 94720, USA. Tel.: +1 5106433726, fax: +1 5106421490.
E-mail addresses: pperera@berkeley.edu, pnpperera@gmail.com (P.N. Perera).

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.06.021

bonds, change of three-dimensional configuration and possible chemical alteration during extraction.
(3) Owing to similarities of monomers, cellulose and hemicelluloses have similar spectral features and therefore high spectral overlap.
(4) Pure components could be spatially collinear at the submicron to micron level, providing identification of them as independent entities is difficult.
(5) Auto-fluorescence from impurities as well as components such as lignin contributes to the noise and shape of the background significantly in the case of Raman imaging.
(6) Raman spectra have poor signal-to-noise ratio (S/N) due to practical constraints such as short integration time that is required to avoid thermal and photo damage to the sample.

As a consequence, supervised calibration methods (e.g. partial least squares [11]) and classification methods (e.g. discrimant analysis [12]) are unsuitable for multivariate analysis of cell wall images. Supervised methods are based on *a priori* knowledge (i.e. a "training set" with known mixtures of the components of interest) of the system to build a reliable calibration model. The prediction accuracy of the model depends on the extent of the similarity of the training set to the actual system. The aforementioned reasons make it difficult to build a reasonable training set that captures the essence of the chemical and structural complexity of the system.

Alternatively, blind (unsupervised) data analysis methods need little or no *a priori* information about the system but progress by extracting all necessary information from the system itself. Here we describe a method that combines *k*-means clustering and spectral entropy minimization with Raman microspectroscopy to elucidate compositional and structural details of plant cell walls.

2. Computational methodology

The image analysis procedure consists of three main steps, (1) spectral preprocessing (2) stepwise clustering and (3) estimation of spectral profiles of pure components and their weights. A general discussion of the implementation of the procedure is presented in this section. The process starts with conditioning of the spectral images for the subsequent steps. Generally poor (low signal-to-noise ratio) Raman spectra were first subjected to wavelet transformation to mitigate the noise contribution. Wavelet analysis has been shown to remove noise contribution with minimal deterioration to the signal by decomposing the original spectrum into a high frequency part ("details") and a low frequency part ("approximations") with a preselected wavelet function. The high frequency component contains the noise contribution to the signal while the low frequency component carries the information content. The low frequency component is further decomposed in subsequent steps and this procedure is repeated until an appropriate level of separation between noise and signal has been achieved. The noise-free signal is constructed by mixing appropriate amounts of high and low frequency coefficients at each level. A detailed discussion of noise removal with wavelet decomposition can be found in the following references [13,14]. Denoised spectra were then converted into their second derivatives with the Savitzky–Golay second derivative (SGSD) method to remove the fluorescence contribution. The second derivative of a spectrum diminishes the slow changing fluorescence contribution while preserving the signatures of Raman peaks [15]. The Savitzky–Golay procedure itself has been shown to remove noise (by means of polynomial fitting) while calculating second derivatives. However, effective removal of noise requires applying larger filter windows which in turn would broaden the Raman features and result in the loss of finer spectral information content. By contrast, noise filtering with wavelet prior to the SGSD gives second derivative spectra with low noise while retaining finer Raman features. Next, the second derivative spectral matrix was decomposed with principal component analysis (PCA) to compress the data matrix into a much smaller size. PCA further removes noise and other undeterministic contributions while conserving the systematic variance in the system [16,17]. The reduced PCA matrix was then introduced to the second step, the stepwise clustering with *k*-means clustering. *k*-Means cluster analysis provides an unsupervised approach to cluster objects into different groups, provided that the user defines the desired number of clusters, *N* [18,19]. The *k*-means algorithm then puts *N* centroids in the space represented by the objects and assigns all the objects to the closest centroid. The position of each centroid is recalculated after all the objects are assigned and this step is repeated until centroids no longer move. *k*-Means cluster analysis has been shown to find suboptimal solutions to the partition. This was overcome by repeating the analysis with different starting points and retaining the solution with the lowest sum of squares. At first, the PCA matrix was partitioned into two clusters to separate cell wall spectra from the spectra that arise from the lumen. Next, the second derivative spectra of cell walls were separated from those of lumen and latter set was discarded. The former set was subjected first to PCA decomposition and then *k*-means clustering with the desired number of classes. The cell wall image was reconstructed at this stage with cluster identities (instead of peak intensity or peak area) to identify and visualize the natural spatial segregation of the image. This stepwise clustering enables the identification of different sublayers in the cell walls of the images. The final step involves identifying and quantifying the factors that contribute to the partitioning of the image. Spectral identities of the clusters were calculated by averaging all the spectra of each cluster. These average spectra were subsequently introduced into the final step of pure spectral estimation.

Spectral entropy minimization methodology [20,21] can be used to recover pure components spectra without assuming their functional forms. The main advantage of the entropy minimization method is that no knowledge of the exact number of pure components and/or estimation of their spectral weights is required. Both these parameters are difficult to obtain for cell wall images. The entropy minimization method relies on the fact that pure spectra have lower entropies (as defined later) than their mixtures. Therefore, the entropy minimization method aims at finding the spectral profile that has the smallest entropy in space that is defined by the mixture spectra. A brief discussion of the entropy minimization procedure is given below. First, the average spectra matrix ($D_{k\times\nu}$, where k is the number of average spectra and ν is the number of wavelengths) was decomposed with singular value decomposition (SVD) [22] as shown in Eq. (1).

$$D_{k\times\nu} = U_{k\times k} \times S_{k\times\nu} \times V^T_{\nu\times\nu} \quad (1)$$

Next, $V_{\nu\times\nu}$ was reduced to retain only physically meaningful j row vectors ($V_{j\times\nu}$ where $j \leq k \ll \nu$. An important attribute of rows of $V_{j\times\nu}$ is that different linear combinations of those can recreate all the mixture spectra in the system as well as the spectra of the pure components. Therefore, the entropy minimization technique pursues the weighting coefficients vector that multiplies the rows of $V_{j\times\nu}$ to obtain the best estimates of pure component spectra that have lower entropies than their mixtures. Accordingly, the first pure spectral estimate ($\hat{a}_{1\times\nu}$) is calculated by a linear combination of j vectors of $V_{j\times\nu}$ weighted by j random numbers ($T_{1\times j}$) between −1 and 1 as shown in Eq. (2).

$$\hat{a}_{1\times\nu} = T_{1\times j} \times V^T_{j\times\nu} \quad (2)$$

$\hat{a}_{1\times\nu}$ was then minimized and refined against a spectral entropy based objective function Obj (Eq. (3)). The first term of the Obj is the

Shannon entropy where h_v is a probability distribution as defined in Eq. (4).

$$\text{Obj} = -\Sigma_v h_v \ln h_v + p \quad (3)$$

$$h_v = \frac{\left|d(\hat{a}_v)/dv^n\right|}{\Sigma\left|d(\hat{a}_v)/dv^n\right|} \quad (4)$$

$$p(\hat{a}_{1\times v}, \hat{c}_{k\times 1}) = F(\hat{a}_{1\times v}) + F(\hat{c}_{k\times 1}) \quad (5)$$

In general, first, second or fourth derivatives (i.e. $n = 1, 2$ or 4) are used to calculate h_v. The second term P is a penalty function (Eq. (5)) that accounts for constraints such as non-negative spectral weights ($F(\hat{c}_{k\times 1})$) and non-negative intensity of the spectra ($F(\hat{a}_{1\times v})$).

The optimum solution is found by checking the value of the Obj function against a preselected threshold and if the criteria have not met, a new $T_{1\times j}$ vector is generated using an optimization criterion such as stimulated annealing (SA) until optimization converges.

Simulated annealing is a stochastic domain search technique that has been proven to find the global minimum in rough terrain where derivative based methods fail [23,24]. SA is inspired by the natural cooling (annealing) process in nature where slow cooling relaxes the system to the lowest energy state while faster cooling traps the system in a higher entropy configuration. SA starts searching the domain at a random point and navigates through the terrain making both downhill and uphill (with acceptance probability less than 1) movements. Temperature is reset periodically and the system is allowed to cool again. These allowed uphill moves and re-annealing help the algorithm to escape local minima and reach the global solution.

Pure spectral estimation of multiple components in a system can be achieved by different approaches. A discussion of these methods and further discussion of the implementation of spectral entropy minimization methodology can be found in the literature [25–27]. We have recovered different pure spectral estimates as follows. The first pure spectrum was recovered by minimizing the entropy of the entire spectrum. Successive pure spectral estimates were obtained by focusing on different regions of the spectrum and therefore avoiding the necessity to subtract a pure spectral estimate from the mixture spectra before recovering the next pure spectral estimate which can be cumbersome for highly overlapped component mixtures with significant fluorescence contribution.

3. Materials and method

3.1. Samples and sample preparation

Air-dried woody samples came from eight-month-old, greenhouse grown *Populus trichocarpa* (Black cottonwood). 50 μm thick cross-sections of hydrated stems were cut with a microtome (Leica RM2265) and placed in D_2O on a microscope glass slide and covered with a glass cover slip (170 μm thick). The edges of the cover slip were sealed onto the glass slides to prevent evaporation of D_2O.3.2

3.2. Data collection

Two-dimensional spectral maps (40 μm × 40 μm) were acquired with a confocal Raman microscope (WITec, alpha300 S, fiber/pinhole diameter = 100 μm), which is equipped with a piezoelectric scan stage. A 100× oil immersion microscope objective (Nikon, NA = 1.40, WD = 0.13 mm) and a laser in the visible wavelength range (λ = 532 nm) were used in the measurements. The linearly polarized laser light was focused with a nearly diffraction-limited spot size onto the samples and the Raman light was detected by a CCD camera (Andor, DV401-BV) behind a grating (600 grooves mm^{-1}) spectrometer (WITec, UHTS 300) with a spectral resolution of ~4 cm^{-1}. The laser power on the samples was approximately 10 mW. The lateral resolution of the system was determined via a knife-edge measurement within the sample fluid cell to be 300 nm, which is near the theoretical limit ($0.61\lambda/NA \approx 230$ nm). Sample areas of interest were mapped by raster scanning in 200-nm steps with an integration time of 200 ms per spectrum, resulting in 40,000 point spectra per image. Results for each sample were obtained in triplicate to ensure reproducibility.

3.3. Data analysis

All calculations were executed with built-in and home-written codes in MatLab 7.9 platform. A sub-spectral region of the 40,000 noisy spectra that span from 750 cm^{-1} to 3300 cm^{-1} was selected for the data analysis. This data matrix was denoised with a Daubechies family level six ('db6') wavelet at decomposition level two with level dependent noise filter. Denoised spectra were then converted into their normalized second derivative spectra with Savitzky–Golay second derivative filter with 21-pixel (~80 cm^{-1}) window to remove the fluorescence contribution. Second derivative spectra were then decomposed with PCA to compress the data matrix and three principal components were retained based on their captured variance values and the shape of the loading vectors. Total variance captured by the PCA analysis was between 92% and 95% for each data set. The PCA scores matrix was then clustered into two groups with *k*-means clustering based on the cosine angle between spectra to identify lumen spectra from the rest and lumen spectra were removed from further analysis. Second derivative spectra of the cell wall spectra were used for PCA (with 10 PCs) decomposition again but only the fingerprint region of 750–1800 cm^{-1} was used. The new PCA matrix was used for *k*-means clustering with ten clusters. The *k*-means clustering step was repeated for 1000 times with different initial centroids to avoid suboptimal partition and the solution with lowest sum of squares error was retained. Spectral identities of the clusters were found by calculating the average spectra. These average spectra were further refined by removing fluorescence contribution by fitting the baseline into a quadratic function. A lower order polynomial was chosen to avoid over-fitting and hence over-subtraction of the baseline. The resultant spectra were used for spectral entropy minimization to find the underlying simpler patterns with simulated annealing with an objective function defined in Eq. (3). The initial temperature in the SA algorithm (Genetic Algorithms and Direct Search Toolbox of MatLab 7.9) was set to 10, the annealing function was set to fast annealing and temperature was updated exponentially. The first spectral recovery (lignin spectrum) was achieved minimizing the entropy of the entire spectral region from 750 cm^{-1} to 1800 cm^{-1}. The second spectrum (carbohydrate spectrum) was recovered by minimizing the entropy in the region from 750 cm^{-1} to 1400 cm^{-1}.

4. Results and discussion

Fig. 1a shows a chemical image of xylem cells of wild type *P. trichocarpa* which has been constructed by integrating the CH stretching region (with WITec 1.94 software) at each pixel. Fig. 1b shows a selected area (as shown by the rectangle in Fig. 1a) of the reconstructed image with 10 clusters after the first two steps (i.e. spectral preprocessing and stepwise clustering) of the procedure described in the Section 2. There are eight sublayers of the cell wall structure including the region of the lumen in the selected area of Fig. 1b (the other two clusters occurred outside the selected area of Fig. 1b). It should be noted here that no spatial information were used during the data analysis and therefore the systematic and symmetrical spatial arrangement of the clusters is self-confirmatory of the fact that our procedure is indeed finding natural

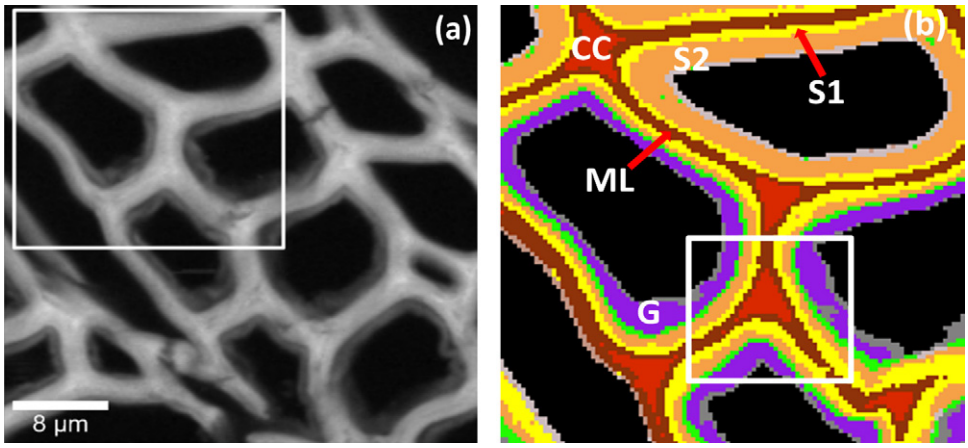

Fig. 1. Confocal Raman image of xylem cells of wild type *P. trichocarpa*: (a) CH intensity image (WITec 1.94 software) and (b) reconstructed image of the marked rectangle in (a) with 10 clusters. Colors in (b) are arbitrary and provide visual contrast of the clusters. Most significant layers of the cell wall are indicated (CC – cell corner, ML – middle lamella, S1 – S1 layer, S2 – S2 layer, G – G layer).

compositional segregations in the cell wall. The number of clusters can be increased or decreased to construct finer or coarser images as desired. However, the maximum number of clusters would be ultimately determined by the S/N ratio of the spectra and/or the number of natural clusters in the image. The reconstructed image displays different layered structures for individual cells. Three cells in the lower half of Fig. 1b have two additional layers (green and blue) which are absent in the top cell. Spectral identities (average spectra) of these clusters are shown in Fig. 2. Although it is possible to work with individual spectra at each pixel to estimate pure spectral profiles, provided that computer power is sufficient, the S/N enhancement achieved by averaging thousands of similar spectra together brings out the smaller features above the noise floor that in turn will yield better convergence of the optimization algorithm. It is also noteworthy that cell wall images are expected to be spatially sparse (i.e. repetition of similar information in space or spatial redundancy) and therefore analysis of individual spectra at each pixel instead of average spectra of clusters may not reveal any additional information but requires much longer computational time. A quick inspection of the average spectra shows significant changes in the relative intensity of the peak at ~1600 cm^{-1} compared to peaks at 1090–1130 cm^{-1}. Two of the spectra that have been recovered from the spectral entropy minimization procedure are shown in Fig. 3 and peak assignments of those recovered spectra are shown in Table 1 in the Supporting Information. Indeed, the assignment of the peaks in the recovered spectra to known vibrational modes of extracted and/or pure lignin and cellulose spectra in literature [28–31] confirms that our procedure has reliably estimated those spectra. One spectrum (red in Fig. 3) carries the characteristic 1590–1670 cm^{-1} peak envelope which has been routinely used to quantify the lignin contribution and the rest of the spectrum agrees well with the reported vibrational modes of the isolated pure lignin spectra reported in the literature. The recovery of other major peaks of lignin (e.g. 1464 cm^{-1}, 1332–1380 cm^{-1} (multiple peaks), 1277 cm^{-1} and 1150 cm^{-1}) allows us to detect monolignol (monomer units of lignin) composition of native lignin *in situ* without ambiguity which is not possible with traditional data

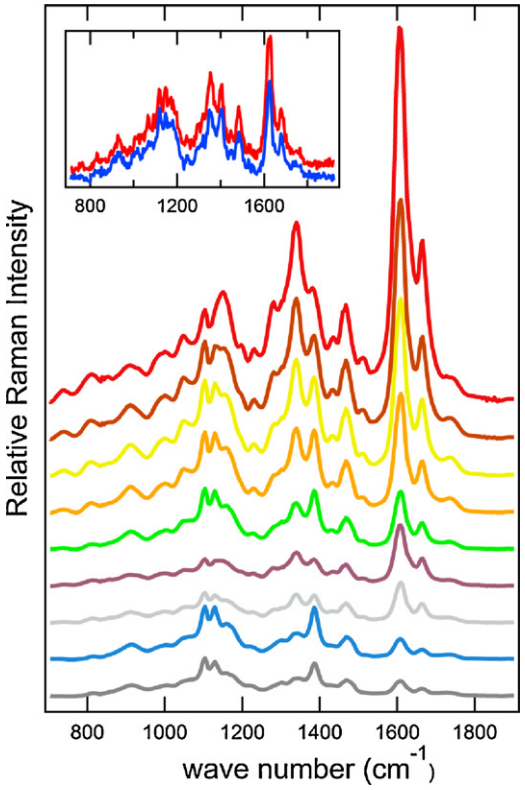

Fig. 2. Average cluster spectra (lumen excluded). Spectra were vertically offset for clarity. Colors of the spectra correspond to different cell wall layers with the same color in Fig. 1b. Inset: Two representative original Raman spectra of the cell wall with significant noise contribution.

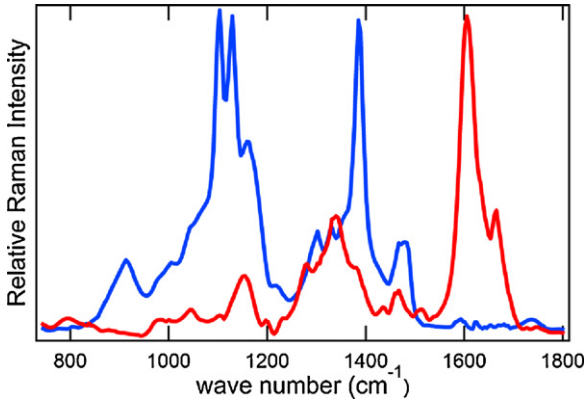

Fig. 3. Two pure spectral estimates recovered by spectral entropy minimization (Red – estimated lignin spectrum, Blue – estimated carbohydrate spectrum). Spectra have been normalized to one at the tallest peak for comparison. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

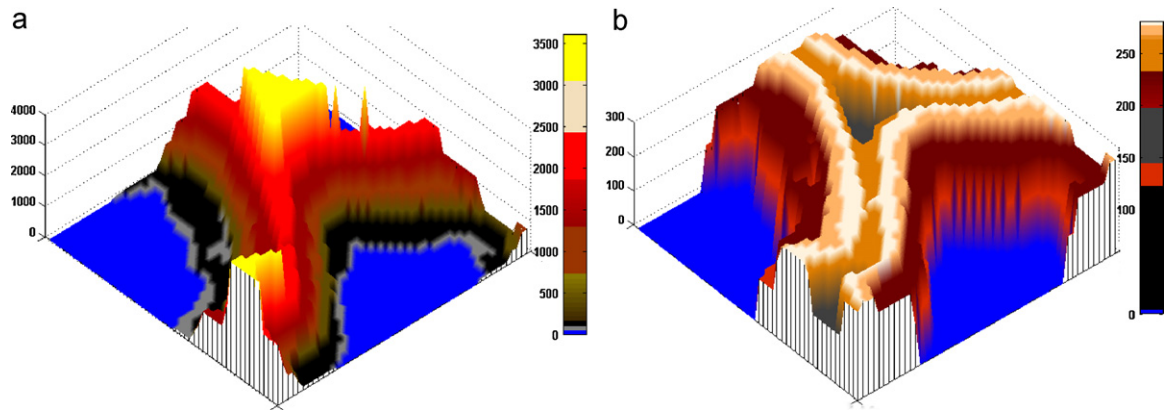

Fig. 4. Relative lignin (left) and carbohydrate (right) distributions across the cell wall as marked (white box) in Fig. 1b. Lignin content differs by as much as a factor of 36 across the cell wall. Carbohydrate content changes by approximately factor of two across major part of the cell wall (see Fig. S1 in supporting information for further information).

analysis methods because of significant spectral overlap with carbohydrate signals. Lignin composition/content plays an important role in biomass degradation as it has been found to be related to the cellulose sachcharification yield. Producing plant phenotypes with altered lignin content/composition is therefore regarded as a viable path to overcome cell wall recalcitrance [32,33]. We have demonstrated the ability of our method to determine and distinguish native lignin composition of different plant species *in situ* without physical extraction in a separate communication [34]. In particular, we determined the composition/chemistry of native lignin of poplar, Arabidopsis and Miscanthus as well as a change in lignin composition upon transgenic suppression of 4-coumarate-CoA ligase (Ptr4CL3) in *P. trichocarpa*.

The second retrieved estimated pure spectrum has characteristic peaks of carbohydrates, both cellulose and hemicelluloses. Spatial segregation or spatially distinct distribution patterns of carbohydrates and lignin across the cell walls allows us to recover pure lignin spectra without spectral contamination from the carbohydrate Raman signals, rendering the structural and quantitative analysis of lignin relatively straightforward. Unfortunately, this is not the case for the carbohydrate analysis. The spatial correlation of cellulose with hemicelluloses makes the spectral recovery of either component in pure form practically unfeasible without additional extrinsic information. The carbohydrate analysis is further complicated due to the fact that Raman spectra of cellulose and hemicelluloses bear significant similarities. However, the recovery of the carbohydrate spectrum free of lignin contribution allows us to quantify and analyze cellulose and hemicelluloses contribution more accurately with extrinsic information on those polymers.

The relative spatial distributions of lignin and carbohydrates were calculated by direct subtraction of first lignin and then carbohydrates from average spectra of the clusters (Fig. 4). The lignin distribution varies (following an approximate Gaussian distribution) across the cell wall by as much of a factor of 36 (and approximately by factor of 7 for the top cells with no additional sublayers) while showing highest accumulation in the cell corners (red cluster in Fig. 1b). The carbohydrate distribution is relatively less marked and varies by a factor of ~2 across the wall and it is relatively invariant through the major part of the wall and is reduced by about ~40% in the cell corners.

The non-invasive, label-free, extraction-free determination of the Raman spectra of lignin and carbohydrates yields a rapid and robust method to analyze the plant cell wall, unlocking structural and compositional characteristics at sub-cellular levels. Cellular specificity can be used to evaluate different types of cells and tissues as well as the cell response to biotic and abiotic stress. This is in contrast to wet chemical and spectroscopic methods currently employed in cell wall analysis that often provide bulk information only, thus losing variabilities within and across different cell types. Carbohydrate spectra can be further decomposed into cellulose and hemicelluloses spectra with use of extrinsic information.

5. Conclusion

In conclusion, our combined Raman microspectroscopy and chemometrics analysis has demonstrated the ability to find spatially resolved compositional information of the plant cell wall. The spectral minimization procedure was used to recover two spectra that represent the lignin content and the total carbohydrate content. The ability to recover entire spectrum of lignin which carries vibratiobnal signatures of monolignol units (e.g. Syringyl and Guacial units) would enable us to extend this procedure to infer structural information of lignin *in situ* without chemical or mechanical breakdown of the cell wall. The presented approach is general such that it can be combined with any multivariate microscopic method (e.g. IR, mass spectrometry, fluorescence, etc.) to analyze any natural or artificial image.

Acknowledgments

We thank Prof. Vincent L. Chiang for poplar samples. This work was supported by the Energy Biosciences Institute, University of California, Berkeley, CA. Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract No. DE-AC02-05CH1123.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.aca.2011.06.021.

References

[1] C. Somerville, S. Bauer, G. Brininstool, M. Facette, T. Hamann, J. Milne, E. Osborne, A. Paredez, S. Persson, T. Raab, S. Vorwerk, H. Youngs, Science 306 (2004) 2206.
[2] A. Carroll, C. Somerville, Annual Review of Plant Biology 60 (2009) 165.
[3] R.H. Atalla, U.P. Agarwal, Science 227 (1985) 636.
[4] U.P. Agarwal, Planta 224 (2006) 1141.
[5] D.S. Himmelsbach, S. Khahili, D.E. Akin, Vibrational Spectroscopy 19 (1999) 361.
[6] N. Gierlinger, M. Schwanninger, Plant Physiology 140 (2006) 1246.
[7] L.Q. Chu, R. Masyuko, J.V. Sweedler, P.W. Bohn, Bioresource Technology 101 (2010) 4919.
[8] B.G. Saar, Y.N. Zeng, C.W. Freudiger, Y.S. Liu, M.E. Himmel, X.S. Xie, S.Y. Ding, Angewandte Chemie (International ed. in English) 49 (2010) 5476.
[9] M. Schmidt, A.M. Schwartzberg, A. Carroll, A. Chaibang, P.D. Adams, P.J. Schuck, Biochemical and Biophysical Research Communications 395 (2010) 521.

[10] N. Gierlinger, L. Sapei, O. Paris, Planta 227 (2008) 969.
[11] P. Geladi, B.R. Kowalski, Analytica Chimica Acta 185 (1986) 1.
[12] J.S. Mattson, C.S. Mattson, M.J. Spencer, F.W. Spencer, Analytical Chemistry 49 (1977) 500.
[13] P.M. Ramos, I. Ruisanchez, Journal of Raman Spectroscopy 36 (2005) 848.
[14] F. Ehrentreich, Analytical and Bioanalytical Chemistry 372 (2002) 115.
[15] A. Savitzky, M.J.E. Golay, Analytical Chemistry 36 (1964) 1627.
[16] S. Wold, K. Esbensen, P. Geladi, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37.
[17] B.C. Moore, Ieee Transactions on Automatic Control 26 (1981) 17.
[18] J.A. Hartigan, Clustering Algorithms , John Wiley & Sons, New York-London-Sydney, 1975.
[19] G.A.F. Serber, Multivariate Observations , Wiley, New York, 1984.
[20] E. Widjaja, C.Z. Li, W. Chew, M. Garland, Analytical Chemistry 75 (2003) 4499.
[21] E. Widjaja, M. Garland, Analytical Chemistry 80 (2008) 729.
[22] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B , Elsevier, 2003.
[23] A. Corana, M. Marchesi, C. Martini, S. Ridella, ACM Transactions on Mathematical Software 13 (1987) 262.
[24] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220 (1983) 671.
[25] H.J. Zhang, W. Chew, M. Garland, Applied Spectroscopy 61 (2007) 1366.
[26] S.T. Tan, H.H. Zhu, W. Chew, Analytica Chimica Acta 639 (2009) 29.
[27] L.F. Guo, F. Kooli, M. Garland, Analytica Chimica Acta 517 (2004) 229.
[28] K.L. Larsen, S. Barsberg, Journal of Physical Chemistry B 114 (2010) 8009.
[29] U.P. Agarwal, S.A. Ralph, Applied Spectroscopy 51 (1997) 1648.
[30] A.M. Saariaho, A.S. Jaaskelainen, M. Nuopponen, T. Vuorinen, Applied Spectroscopy 57 (2003) 58.
[31] K.K. Pandey, T. Vuorinen, Holzforschung 62 (2008) 183.
[32] W.J. Hu, S.A. Harding, J. Lung, J.L. Popko, J. Ralph, D.D. Stokke, C.J. Tsai, V.L. Chiang, Nature Biotechnology 17 (1999) 808.
[33] F. Chen, R.A. Dixon, Nature Biotechnology 25 (2007) 759.
[34] P.N. Perera, M. Schmidt, V.L. Chiang, P.J. Schuck, P. D. Adams, Anal. Bioanal. Chem, submitted for publication.